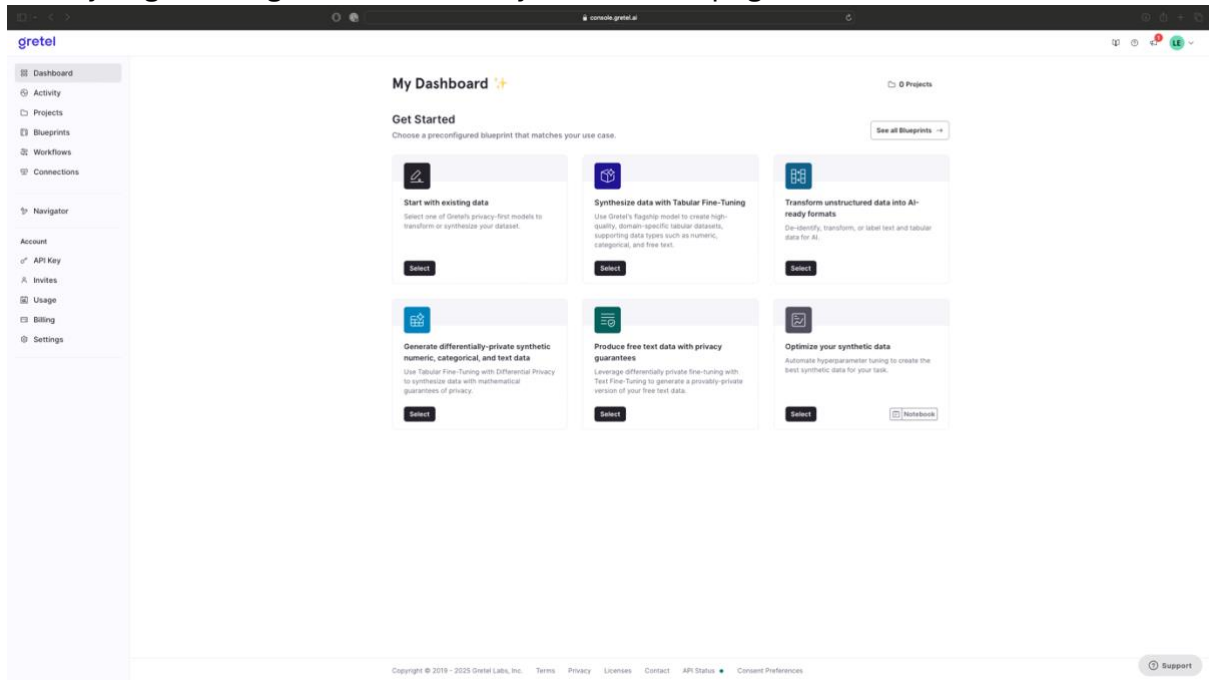


This document will provide a quick overview of using Gretel.ai for synthetic data generation.

First off, go to this link. <https://gretel.ai>

It asks for a work email address but I had success signing in with my personal gmail account and my deakin student email address.

Once you get through the verification you'll see this page.



Click on Start with Existing data and you'll be prompted to create a project.

Dashboard / Start with existing data ⓘ



Step 1 - Project
Current

Step 2 - Model
Upcoming

Step 3 - Input data
Upcoming

Step 4 - Configuration
Upcoming

Choose a project

Select an existing project or create a new one

Select a project

Create a new project

Project name

synthetic data creation

Continue

Cancel

Select your model. Have a read of the descriptions but the top one that is preselected appears to be the most relevant.






[Dashboard](#) / [Start with existing data](#) ⓘ



Step 1 - Project
synthetic da...ation > Step 2 - Model
Current > Step 3 - Input data
Upcoming > Step 4 - Configuration
Upcoming

Select a Model

Choose a Gretel model for your use case

-  **Tabular Fine-Tuning**
Use our flagship model to generate privacy-preserving synthetic data across categorical, numeric, time-sequence, and text fields. Selected
-  **Text Fine-Tuning**
Fine-tune large language models with differential privacy, evaluate model quality, and generate synthetic text while protecting sensitive training data.
-  **Transform**
Detect and transform sensitive data with configurable templates, validate data quality, and automate PII discovery across your structured data.
-  **Tabular GAN**
Quickly generate synthetic tabular data for high-dimensional datasets while preserving relationships between numeric and categorical columns.
-  **Tabular DP**
Create privacy-protected synthetic data with mathematical guarantees using fast graph-based modeling optimized for tabular structures.

Continue

Cancel

The next page is where you drop in your input data that you want the model to be trained on. I think this will be extremely useful for creating transactional data. Just make sure we're covering the categories agreed on in the meeting last night.

[Dashboard](#) / [Start with existing data](#) ⓘ



Step 1 - Project
synthetic da...ation > Step 2 - Model
Tabular Fine-Tuning > Step 3 - Input data
Current > Step 4 - Configuration
Upcoming

Where's your data artifact?

Let's begin by defining your input data

- ☒ **I have a data artifact ready to upload.**
Click the button to select a local file, or drag and drop it into the upload window.

Choose file

or drag and drop a CSV, JSON(L), or Parquet file here to upload


- ☐ **I'd like to connect to my external data source.** New
Select an existing connection or create a new one.

- ☐ **I don't have a data source.**
No worries, just use our sample dataset to get started.

Continue

Cancel

You'll then be at the final page before running the process. It will be automatically set to generate 5000 records, if you want to change this scroll down and edit the yml file.

[Dashboard](#) / [Start with existing data](#) 



Step 1 - Project
synthetic da...ation

Step 2 - Model
Tabular Fine-Tuning

Step 3 - Input data
Wed04Sep...ths 1.csv

Step 4 - Configuration
Current

Almost done!


We've chosen a model configuration based on your selections

Gretel configuration 

navigator-ft

Recommended

or upload your own

 Upload your own no file selected
Must be .yaml or .yml

navigator-ft.yml

Edit

```
14 ..... # This is useful if your records are sequential.
15 ..... # Note that this parameter can only be used when
16 ..... # your records are grouped using the above parameter.
17 ..... order_training_examples_by: null
18
19 ..... generate:
20 .....   num_records: 5000
21
22 ..... params:
23 .....   # The parameter below is a proxy for training time.
24 .....   # If set to 'auto', we will automatically choose an
25 .....   # appropriate value. An integer value will set the
26 .....   # number of records from the input dataset that the
27 .....   # model will see during training. It can be smaller
28 .....   # (we downsample), larger (we resample), or the same
29 .....   # size as your input dataset. A starting value to
30 .....   # experiment with is 25,000.
31 .....   num_input_records_to_sample: auto
```

Useful Links

- [Need help? Check out our configuration docs.](#)
- [Tips to improve model quality and accuracy.](#)
- [Learn more on privacy protection.](#)
- [Have questions? Join our Discord community today.](#)

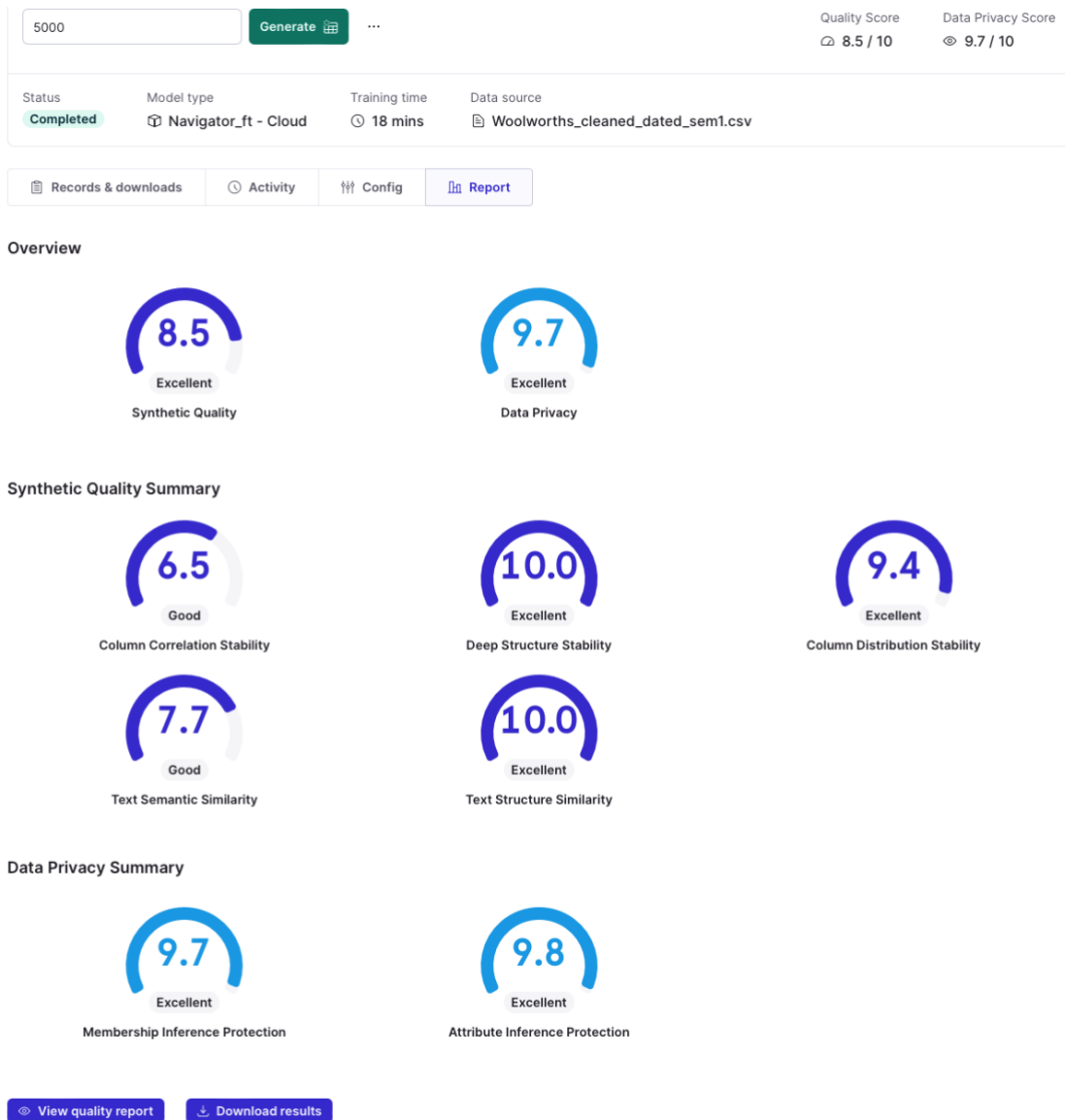
YAML

Run

Cancel

Once you start the process it can take a while to complete. Some of my attempts took between 10 and 20 minutes. (Depends on amount of data you use)

You'll receive a report on the quality of the data along with the download button for the data you just generated.



It is possible to do this through terminal using the CLI and the Gretel API key which is available, but the web process is much more straightforward for the same results.

Here is a link to documentation. <https://docs.gretel.ai>

In my experimenting I combined the Woolworths csv files into one dataframe and added a date column. Gretel did not come up with new dates but instead generated for those dates. So, if you were to work on grocery price generation another approach would be needed.

Making me lean to focusing on using it for transactional synthetic data.